DOCUMENT RESUME

ED 148 897 TM 006 838

AUTHOR O'Reilly, Robert P.; Schuder, R. T.
TITLE Some Issues in the Measurement of Basic Competence in Reading.
PUB DATE Apr 77
NOTE 51p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS Basic Skills; *Cloze Procedure; Criterion Referenced Tests; Diagnostic Tests; Elementary Secondary Education; Measurement Techniques; Multiple Choice Tests; *Reading Ability; *Reading Comprehension; *Reading Tests; *Test Construction; Testing Programs; Test Reliability; Test Validity
IDENTIFIERS *Competency Tests; Domain Referenced Tests; *Multiple Choice Cloze (O Reilly and Schuder)

ABSTRACT

The content and utility of two models of reading comprehension tests, the basic skills assessment model and the multiple choice cloze (MCC) model, are discussed. The basic reading competency model assesses the ability to read or infer "basic" meanings from a range of written or symbolic communication, as is necessary for daily living in this society. This limited analysis of the basic competency model suggests some degree of positive practical impact of the test; applications of this model are illustrated. Criticisms of this model state that such a test is not theoretically based, nor objectively reproducible, nor does it yield scores that provide detailed diagnostic guidance in the instructional process. The multiple choice cloze (MCC) technique was developed as one expression of the literal comprehension construct, viewed as the basic skill underlying the reader's ability to infer explicit meanings from written language. Items in the test are objectively developed and derived from a theoretical framework, thus providing guidelines for interpreting the resulting test scores. Current developments suggest that these test results can be used to identify the kinds of materials that readers can comprehend at the literal level for specific purposes. The test also appears to offer a diagnostic capability with fairly clear-cut instructional implications. (Author/MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED148897

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Robert P. O'Reilly

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Some Issues in the Measurement of Basic Competence in Reading

By

Robert P. O'Reilly and R. T. Schuder
Montgomery County Public Schools*
Rockville, Maryland

TM006 838

*Paper read at the Annual Meeting of the American Educational Research Association, New York City, April, 1977.

The measurement of basic competence in the three R's seems to be the latest movement in the continuing preoccupation with the output of the public schools, with reading perhaps being of greatest concern. This concern issues from a gradual and general eroding of public and governmental confidence in the nation's schools, based in part on awareness of sharply escalating educational costs together with paradoxical indications of less than satisfactory or reduced qualitative output. The wide-ranging ten-year decline in standardized achievement test scores is a case in point, although it is not at all clear where the responsibility of the schools lies in this phenomenon (Harnischfeger and Wiley 1976). Nevertheless, it appears the schools have increasingly become the target of legislation or other governmental control designed to insure that all students achieve a basic minimum of competence in reading, writing, and arithmetic skills.

Drawing inspiration from research dealing with the definition and measurement of basic competence in adults, such as the Harris Survey (1970), the Educational Testing Service (ETS) study of adult functional literacy (Murphy 1973), and the Adult Performance Level Study (Northcutt, 1975), the basic skills issue has spawned a new type of test, one that differs in significant ways from the tests used heretofore to measure basic competence. Examples of this new type of test are the New York State ninth grade basic competency test which drew on the adult reading tasks developed in the ETS study, the Maryland Functional Reading Test which presents the testee with a set of real-life reading tasks, the new ETS basic skills tests scheduled for release in May which will

apparently draw on the adult literacy study for its reading items, and the Madison Public Schools literacy tests assembled by Hansen and Hesse (1974).[1]

The ETS and other basic competency tests developed (or under development) in the reading area represent in some respects a significant departure from the tests previously used to assess basic skills, namely the familiar batteries of standardized tests used in the schools. As the pattern of this development has now clearly emerged in the reading area, it becomes possible to evaluate its potential impact on the educational process. One may reasonably ask how the new basic competency tests differ from the tests the schools have previously relied upon, and whether they represent a useful adjunct to the decision-making needs of school personnel. Or, alternatively, does the new testing methodology simply represent another administrative burden that will only add another layer of questionable information that few, if any, decision-makers use? It is the purpose of this paper to address these and other issues in the limited context of comprehension testing--a major focus of the new basic skills assessment model.

The term basic skills assessment model is used here to refer to the archetype which has evolved for assembling the comprehension sections of a number of competency tests which seem to have converged in a common methodology. This model is the subject of the present analysis which begins with a simplified outline of its characteristics and proceeds to

[1]Several additional tests of basic reading competence for adults were recently reviewed by Nafziger, Thompson, Hiscox, and Owen (1976).

evaluate both its negative and positive features as they are relevant to deciding the model's utility in the schools. In a second section of the paper, an alternative model is offered as a more advanced approach to the problem of assessing basic comprehension skills, one that seems to go further in providing the schools with a methodology that will provide useful data for educational decision making.

A Model for Testing Basic Comprehension Skills

The evident purpose of the extant model of basic reading competency is to assess the testee's ability to read or infer the "basic" meaning in a range of what are loosely referred to as written and/or symbolic communications necessary to some modicum of survival in American society. To that end, a variety of common stimulus materials presenting symbols, pictures, words, lists of brief phrases or statements, continuous written discourse, maps, directions, or some combination of these or additional stimulus categories are assembled into a test. A series of questions are then asked on each stimulus to determine the testee's ability to effectively respond in what amounts to a simulated context of a real life "reading task." The testee's responses are then summed across categories of stimulus materials to provide a set of part scores and a total score. The Maryland Functional Reading Test (MFRT) given at grades 7 and 11, for example, presents the testee with a series of multiple-choice tasks grouped into categories: locate references, understand forms, gain information, and follow directions. A composite of the stages in developing the MFRT and other tests are given in the following sections.

Ecological Validity

The basic competency test is distinguished in part by a generally more extensive and systematic attempt to represent in the test those reading tasks which are relevant to the testee audience than has been the case with the typical norm-referenced test of reading. In the initial stages of test development, there is some attempt to define the audiences, their characteristics, and their needs and activities in a variety of important reading situations. The importance of reading tasks or situations is defined in terms of some mix of economic, social, political, cultural, and academic benefits that may accrue to the reader or his society by virtue of his being able to respond effectively in situations requiring reading and other skills.

The ETS study of adult functional reading competence carefully categorized the activities and reading materials of a nationally representative sample aged 16 years and older and further tabulated the results in terms of the respondents' sex, income, employment status, and judgments of the importance of the material, and so on. Northcutt in the Adult Performance Level Study (APLS) used an even more involved approach to the importance issue by selecting reading tasks that discriminated between successful and unsuccessful adults. In the test development effort undertaken in the Madison, Wisconsin Public Schools, Hansen and Hesse (1974) used panels of teachers and parents to identify the reading materials that were important in the lives of students in grades 1-12. This activity produced a scheme for categorizing reading materials and tasks in 12 major categories, each of which contained as many as 10 subdivisions.

Construction of Reading Tasks

The identification of relevant stimulus or task domains in the basic competency model provides a systematic basis for choosing source materials for the test. Some 250 different reading tasks were initially assembled in the ETS study from those categories of reading materials judged to have some importance in one or more significant adult life roles or activities. The New York State Education Department sampled from some of the ETS categories to derive its 9th grade basic competency test in reading. Hansen and Hesse used their categorical schema to sample 300 reading passages and then assembled these into 84 test forms for use in grades 1-12.

The conventional test development methodology, by contrast, resulted in a severely limited and biased sampling of passage content. (Level III of the California Achievement Test contains only three short passages). The types of items written were rigidly controlled through a traditional reading skills classification which included terms such as main idea, detail, interpretation, inference, etc. Certain of these question writing paradigms resulted in excessively difficult items which were more dependent on the reader's background knowledge and test-taking skills than they were on his processing of the test passages (Tuinman 1973-74). The basic competency test presents the testee with a much wider array of stimuli or exercises (Form B of the MFRT contains 27), identifies the stimuli which appear to require a common type of processing as skills, and limits the types of questions asked to simple inquiries involving interpreting or locating one or more explicit elements in the stimulus.

An informal examination of the content of Form B of the MFRT, using the reading comprehension item types distinguished by Anderson (1972), and Bormuth (1970), indicated that about one-third of the items approximated the verbatim and/or paraphrase type questions, another one-third were a locating information type question in which the testee is required to find the stimulus that contains a verbatim or symobolic transformation or paraphrase of the question, and the remainder of the questions involve distinguishing instances of concepts or principles in the information in the stimulus, as well as some more complex combination of the interpreting, locating, and concept item types. Questions in the test appeared to be answerable from the information given; there appeared to be no instances of the complex inferential questions that characterized tests such as the comprehension sections of the Stanford or Metropolitan Achievement Tests.

Generally, it appears that the comprehension section of the basic competency test is a test of literal comprehension of sorts, although no specific paradigms seem to be evident for controlling the various types of literal items that might be written. A variety of item formats are possible, including the innovative cloze format used by Hansen and Hesse (1974), but it seems likely that the most widely used tests will be of the multiple-choice type. The original item formats developed in the ETS study, for example, appear to be very largely capable of conversion to a multiple-choice format, and as noted previously, this effort is apparently well underway.

Item Selection and Test Development

Consistent with the procedures used for assuring a greater degree of ecological validity in the selection of reading tasks, item revision in the basic competency model generally involves cleaning up the items and attempts to discover various sorts of bias and other types of noise in the items through field trials and expert panel review. Typical item selection criteria used in the development of norm-referenced tests, such as difficulty levels and discrimination indexes, are not used in test development. There seems to be some concern with homogeneity of scores and subscores, but it is not clear whether test reliability coefficients play a role in item selection or whether the appropriate role is in the design of consistent item formats.

Eventually, the items are organized into a test or tests (the MFRT has three forms for different age level populations) with instructions for administration. Judging from the models available to date (Nafzinger et al., 1976), one may expect the comprehension section of the basic competency test to have about 50 items, divided into 4-6 subtests each identified as a skill category representing a class of information or a class of reading tasks. The testee is to be allowed adequate time to attempt every item on the test without pressure.

Standards of Competence

The basic competency model is a domain-referenced test, and a score thus carries the implication that the testee can perform at some level in a universe of similar tasks. This requires that an acceptable standard of performance be set, and the standard is typically an arbitrary

criterion of 80 to 90 percent correct across all items in the test or subtests thereof (Nafzinger et al., 1976). The new ETS tests of basic competency will reportedly use a sliding criterion that can be adjusted to meet local standards. Bormuth's (1971) work with the cloze technique provides one of the few empirically validated performance criteria [2/] for judging basic reading competency, allowing the user to identify whether the stimulus material is: (a) above the testee's level of literacy; (b) suited for his instruction; or (c) can be read by him independently.

Given the nature of the tasks presented to the testee, it may be expected that very high proportions of the testee population will meet the standards of competence on typical tests constructed in accord with the basic competency model. Table 1 confirms these expections with some data from the recent administration of the MFRT to the 7th and 11th grade students in Montgomery County, Maryland. (Students in the 7th grade received an easier form.)

TABLE 1

Average Percent Correct for Montgomery County 7th and 11th Graders on the Subtests for the MFRT in 1976 (N = Approximately 8000/Grade)

	Locate Reference	Understand Forms	Gain Information	Follow Directions	Total
7th Graders	85.0	80.0	80.0	88.0	84.0
11th Graders	91.0	93.7	94.9	96.9	94.2

2 The reference here is only to the fact that this study provides one of the few instances of an attempt to empirically validate passage performance standards, not whether the study actually succeeded in doing so.

Utilization of Normative Data

In addition to flexible and inflexible standards of competence, the basic competency model offers (or will offer) normative data for various populations on the tasks and categories represented in the test, NAEP style. The Maryland State Department of Education (MSDE) provides annual summary data on the MFRT for every school and county as well as the state as a whole. In the ETS study, some 170 tasks were administered to a national sample of 7,866 adults, and it appears there are plans to generate normative data for the tasks to be included in the new competency tests scheduled for this May.

The foregoing analysis suffers from all attempts to simplify and distill, but nevertheless should prove instructive in the present effort to understand the practical and theoretical implications of the basic competency model as it effects the assessment of reading comprehension in the schools. The tests that result from this model are a form of domain-referenced test designed to monitor the achievement of basic reading skills at strategic points in the curriculum, e.g., the 7th and 11th grades. However, the methodology has been extended to the junior high school in some instances and, it can be expected, will generalize to other grade levels. In Maryland, new legislation requires that basic competence in reading be measured at every grade level. New York State has long had a regulation which required the annual measurement of basic competency among some elementary grade populations. The Maryland regulation, however, is more consistent with the basic competency model and far more pervasive in its implications for assessment and instruction. Materials recently distributed to the schools throughout

the state set grade level performance standards in terms of a readability level for each grade and define comprehension largely in terms of performance with different categories of reading materials relevant to the grade 1-12 student population.

Implications of the Basic Competency Model for Comprehension Assessment in the Schools

The basic competency model in reading assessment signals a number of points of departure in educational measurement and appears to be part of a larger innovative thrust that promises to provide some long needed improvement in the meaningfulness of testing in the schools. This larger thrust includes the many new assessment devices represented in the criterion-referenced testing movement, the NAEP model (1972), and the National Institute of Education's recent investments in a wide variety of research and development projects designed to improve the psychological and educational meaningfulness of basic skills testing (1975).

One of the more meaningful departures represented in the new basic competency model in reading is the feature of domain-referencing originally espoused by Hively (1974). Although some important elements of Hively's model are lacking, the principle as applied will tend to produce a test that more systematically represents the task universe of interest. This feature, together with the types of questions written to date, should tend to produce a more valid reading test by reducing the incidence in the test of culturally biased content (Bormuth, 1970), trick questions, questions that are passage independent (Tuinman, 1973-74), and questions that emphasize higher order reasoning as opposed to

processing the meaning of the stimulus per se (Carver, 1973). The broad sampling of stimulus material should also tend to promote recognition of the need for flexible test formats capable of accommodating the large numbers of test forms required to get adequate samples of performance in the content universes of interest.

A second point of departure lies in the improved fit between the characteristics of the testing model and the kinds of educational decisions that are made on the basis of test data. Norm-referenced testing has been frequently criticized as an irrelevant model for the educational decisions it is mostly used to support. What implications are suggested to a parent for example, by knowing his child is two or more years below grade level reading performance in seventh grade? In what situation requiring reading performance is the child unable to perform? The basic competency model, by contrast, presents the testee with a series of straightforward simulations of the context in which the stimulus materials play some meaningful function. Mastery of a category of reading tasks, not comparisons with others who scored on the test, is the issue of interest and, because of the method of selection and presentation of the stimulus material, correct responses would appear to have some utility as predictors of performance in a larger universe of situations. With norm-referenced testing, it is not clear with which stimuli the testee can perform, unless it be another reading test.

Because of its closer coordination with educational decision-making functions, the basic competency model should be more useful in some

respects than conventional test data to parents, teachers, and other decision-makers. As the tests now stand, the data represent a kind of bottom line for the performance of educational systems, and the test results can be further used to identify students who are especially needful of help at one or more critical points in their educational development. In Maryland, the results from the annual administration of the MFRT are published district by district and for the State as a whole in the local newspapers.[3] The effect has been to emphasize instruction in functional reading skills, potentially facilitating the reading performance of those students who might ordinarily maintain the nation's pool of functional illiterates.[4]

A third point of departure lies in the possibilities for developing new knowledge of reading performance as a function of the types of normative data yielded by the basic competency model. Little is known of how reading performance develops in children from the point of view of the cognitive and linguistic demands made upon them by various reading tasks. The reading competency test provides the beginnings of an attempt to sort reading tasks into meaningful groups and to characterize the performance of different populations on these tasks.

The foregoing discussion represents some of the more important and noticeable advantages of the basic competency model. From this brief analysis, it seems fair to conclude that the model in practice offers a degree of improvement over the tests used to assess basic reading

---

3 See, for example, Md. Reading Scores Improving, But Many Students Lack Basics. (The Washington Star, March 15, 1977)

4 Estimated as high as 20% of the adult population by APLS project staff.

skills to date, although it is difficult to see how it offers any theoretical advances. Some major criticisms of extant achievement tests have been given some attention in the model (Bormuth, 1970), and it appears to presage a broader range of advances to come in the development of tests in reading and other areas.

Whatever its advantages for current practice and au:ries for the future, the basic competency model has some serious limitations that are shared with the standardized tests it will increasingly replace. Though the model represents some progress, it has not gone nearly far enough and, consequently, its impact on knowledge and practice in reading will be relatively minor. This is because the basic competency model still remains largely a creature of the test-maker's craft rather than of his theory. Some criticisms in this vein follow:

1. Construct Validity - Achievement test makers rarely heed Messick's (1975) advice on the need for theoretical clarification of what is measured by the test, and basic competency test-makers are no exception. Judging by some reports (Nafzinger et al., 1976; Murphy, 1973; Northcutt, 1975; Bormuth, 1971), there has been a general avoidance of defining the meaning of such related concepts as basic literacy, basic reading skills, and literal comprehension. Instead, the tendency has been to define basic competence in reading circularly, i.e., in terms of performance on the categories of stimuli or tasks displayed in the test, which as noted, are generally chosen in terms of their social utility. The consequences of this failure to provide a basis for investigating the construct

validity of the test are well known. Generally speaking, the users of such tests are not in a good position for interpreting the results of use, nor is the test maker in a good position to guide the development, interpretation, and use of his test in research and practice.

2. Subjectivity of Item Writing - The objective generation of test items in an achievement test largely involves providing for the elimination of bias in the selection of test stimuli, together with rule oriented procedures for the construction of items (Bormuth, 1970). Bias in the selection of test passages or other such stimulus materials is less an issue in the competency model, but little systematic progress appears to have been made in the development of objective procedures for writing items for each stimulus category. Item construction in the model is still a consequence of the item writer's craft, with unknown consequences for the content of the test.

3. Predicting Position in the Stimulus Domain - To date, competency tests are scored in much the same way as traditional tests: responses are summed across an unscaled array of test passages or other displays for each domain of interest and the total test. This is an inadequate structure for a test where the obvious purpose is to determine the kinds and levels of stimulus material the testee can process--the easier passages that occur in the feature pages of newspapers, for example. A more adequate methodology would begin like the basic competency model with

systematically sampling the stimulus domains relevant to the audiences. From there, some methodology, such as a readability formula, is required for scaling materials in the domains and defining the domain distributions of the scaled scores. Subquently, stimulus materials may be sampled from the range of scale values in each domain, converted to suitable item forms, and rescaled for the test. Assuming that the item type and the method for scaling stimulus complexity access the same linguistic features, there should be a high correlation between the two scales thus supporting the ability to predict from a point in the test to a point in a domain distribution. Or, in other words, it becomes possible to accurately predict the types and levels of materials the testee can comprehand or process in a given domain, such as feature stories.

4. Diagnostic Capability - Ascertaining an individual's position in a number of critical domains of reading is a desiderata in a competency test, but it is of limited value in determining the kinds of programmatic actions that might be taken with the testee. However, the model is not well structured for yielding scores with instructional implications, for the items are not designed to reflect the nature of the demands the test stimuli make on the reader. It is difficult to see, for example, how any diagnostic implications can be drawn from the stimulus categories locating information and understanding forms, terms which elucidate the nature of the reader's problem no better than the traditional reading skills, such as reading for detail, interpretations,

generalizations, and so on. Currently, the basic competency model appears to offer the reader only the vague advice that he needs to work on tasks similar to those he failed.

The lack of diagnostic utility is a direct consequence of the failure in the model to grapple with the conceptual, semantic, and syntactic demands made on the reader by the various reading situations of interest. Simulating the superficial outward demands of reading tasks and grouping them into sets based on evident demands of the tasks fails to scratch the surface of the nature of the interaction between the reader and the task. Until the question of the relationship between the reader and task is examined effectively, there can be no serious pretense of meaningful identification of skills involved in the basic competency model.

5. Meaningful Performance Standards - Bormuth's (1971) pioneering research on passage performance criteria demonstrates that the problem of setting meaningful cutoffs or standards of performance for reading is exceedingly complex. The model used in this research, which employed a number of criterion groups, still appears to be the most adequate resolution of the problem of setting interpretable cutoff scores--at least for reading tasks involving connected discourse. Whatever the resolution of this problem, the basic competency model is yet at an early level of development on this issue, for it simply follows standard rule-of-

thumb methods used in criterion-referenced testing. Until this problem is satisfactorily resolved, users of the test remain unable to distinguish programmatically among testees with different scores below the passing point.

This limited anlysis of the advantages and limitations of the basic competency model suggests some degree of positive impact of the test in practice, but this impact will be necessarily limited, for the degree of improvement obtained over conventional testing technology is relatively slight. The new tests of basic competency in reading should prove to be more context sensitive, but it can be seen that scores from the test cannot be used to identify how well the testee will perform at a particular point in the range of situations available in a given domain of reading tasks. A test from the model is not theory based, nor objectively reproducible, nor does it yield scores that provide much in the way of detailed guidance in the instructional process. The model offers the beginnings of a crude taxonomy of reading tasks, but this taxonomy is yet very superficial and fails to relate to the problem of describing in meaningful terms how the various readings tasks of interest differ in terms of their implications for the reader. Little is also known of how the various basic competency tasks relate to school reading tasks, for the latter remain about as ill-defined in their psycholinguistic implications for the reader.

An alternative model under development by the authors over the past three years appears to offer a more comprehensive approach to the problem of accessing basic skills in reading, although the approach

is presently limited to reading situations involving continuous discourse. The next and concluding section of this paper outlines the major features of this model and attempts to show how tests derived from this model more effectively meet the needs of those interested in measuring basic reading skills.

The Multiple-Choice Cloze: A Test of Literal Comprehension

The test development project outlined here began as an attempt to assemble a domain-referenced test of literal comprehension, one that would effectively address deficiencies noted with reading assessment techniques in an extensive review of the literature on reading comprehension measures (O'Reilly, Schuder, Kidder, Salter, and Hayford, 1976). A somewhat later paper (Schuder, Kidder, and O'Reilly, 1976) further outlined the construct of literal comprehension, distinguishing it from more complex levels of comprehension that appear to be dependent on a first processing of text at the literal level. This construct of literal comprehension has considerable overlap with the level of meaning accessed by current tests of basic competence in reading. Whatever the similarities, literal comprehension is viewed as the basic skill underlying the reader's ability to infer the explicit meanings in a wide range of written language units and messages of various lengths and types, ranging from phrases and clauses to sentences and discourse of book length.

A modified version of the cloze technique, referred to as the multiple-choice cloze (MCC) was developed as one expression of the literal comprehension construct and was used to assemble a testing system. The MCC item format, as shown in Figure 1, resembles a standard cloze passage attenuated in length. Passages are generally about 70 - 80 words long (in grades 1 and 2, they range from 25 - 45 words), and they exhibit the standard deletion rate of approximately every fifth word. The item form, in contrast with the standard cloze, is designed

Mother had been more than a little ___27___ with Father on the day that he had ___28___ Muley home. "Another ___29___ !" she exclaimed. "When Betsy and Suzie ___30___ us all the milk we ___31___ ? I do believe, Jonathan, you ___32___ no sense about ___33___ . I wish that you had bought us a ___34___ instead of another cow."

Father ___35___ , "We have one excellent ___36___ in the house. That's enough."

(27) A. expert
B. angry
C. unusual
D. rusty
E. current

(28) A. brought
B. speared
C. mailed
D. floated
E. piloted

(29) A. mist
B. aunt
C. cow
D. upstairs
E. spray

(30) A. give
B. loose
C. pronounced
D. mold
E. weave

(31) A. need
B. shine
C. race
D. print
E. recite

(32) A. tune
B. exercise
C. pump
D. have
E. gobble

(33) A. scrambles
B. limits
C. jackets
D. mills
E. cows

(34) A. star
B. tower
C. clock
D. slab
E. road

(35) A. tested
B. gazed
C. said
D. trailed
E. rushed

(36) A. sheet
B. swim
C. timepiece
D. council
E. skyscraper

Figure 1 A Multiple-Choice Cloze Passage

to access more of the semantic content which is peculiar to a given text; nouns, verbs, adjectives, and adverbs are deleted rather than any-word. This deletion type should result in a more viable measure of sentential and intersentential meaning than the standard cloze, since information necessary to replace content words tends to be distributed throughout the passage (Fillenbaum, 1963). The every-fifth-word deletion pattern is adhered to as much as possible, given the selective deletion type, in order to sample as objectively and thoroughly as possible among the surface representations of the syntactic and semantic relations in the text. The item form is invariant across grades K through 12.

Some 1500 passages ranging in length from approximately 25 to 80 words were sampled from the following domains, representing texts and other written messages that students and adults are expected to read:

1. Textual materials in reading, literature, language arts, social studies, science, and mathematics
2. Citizens materials (newspapers and news magazines)
3. Consumer materials (instructions, catalogs, advertisements, etc.)
4. Reference materials (test instructions, encyclopedias, etc.)

The sample passages were then converted to the MCC format and were scaled on readability level based on a division of the Spache and Dale Chall formulas into 28 equal intervals.

Distractors for each passage were generated by a computer program that randomly accesses sets of words from a 12,400 word vocabulary list

within the constraints of: (a) the grammatical class of the deleted word; (b) whether the deleted word is a "content" word or is a core (basic) vocabulary word; and (c) the grade level of the deleted word. Any given distractor thus functions grammatically but not semantically in the position of the deleted word, is at the same graded reading level as the passage source, and is a content specific word or a core word as required to match the subject area to which the word belongs. The core or general vocabulary lists were compiled from Harris and Jacobson's Basic Elementary Reading Vocabulary (1972) and the EDL Research and Information Bulletin 5: A Revised Case Vocabulary (Taylor, Frackenpohl, and White, 1969). The content-specific word lists were compiled from both the Harris and Jacobson source and the American Heritage Word Frequency Book (Carroll, Davies, and Richman, 1971).

The MCC Format As A Model For Assessing Basic Reading Competence

The MCC item format is an attempt to build a flexible and generalized test assembly resource that would be useful in part in discovering how realistic and objective standards of reading competence might be set for a wide variety of reading situations, including those in the levels of the elementary and secondary curriculum. The basic concepts which underlie the present or projected development of the MCC passage corpus are outlined briefly in Table 2. The first six of these concepts are of particular relevance to the present discussion and are discussed in more detail in the text that follows:

Table 2

Properties and Advantages of Multiple-Choice Cloze Testing Materials

Property	Advantage
Objectivity in item form	The format of the multiple-choice cloze adheres to the concept of an item form, an objective and generative procedure for producing items that are an unbiased representation of a universe of content. The application of this principal here avoids the problems of subjectivity and resultant content bias in test construction procedures, criticisms which have been characteristically leveled at tests of reading comprehension.
Unidimensionality	The multiple-choice cloze format is a unitary and generalizable measure of reading comprehension, appropriately termed "literal comprehension." Literal comprehension is the reading behavior most likely to be affected by the instructional program. It is also the level of comprehension that is important for the maintenance of a reasonable level of social and economic survival.
Passage dependency	The logical requirement that the student's response to the test situation be dependent upon actually reading the test passages is met by the nature of the task, i.e., completing deletions. There are no questions apart from the passages.
Domain-referenced content	This feature of domain referencing, with the application of additional technology, will provide a formal link with local reading materials. Test scores may thus be ultimately used to select reading materials for individuals or groups or to determine the "fit" of available local materials to the reading abilities of certain populations of readers.
Equal-interval scaling	Cloze passages are ultimately to be calibrated on a single equal-interval scale with meaningful upper and lower limits. This represents a substantial improvement over existing scaling procedures in tests of reading comprehension. Any test assembled from the item pool will be referenced to this single scale which in turn can be related meaningfully to performance

(Table continued on next page)

Table 2 (Continued)

Property	Advantage
	criteria identifying passage performance for an individual for any range of readability levels.
Generation of instructionally relevent part scores	The test passages can be used to produce theoretically relevant passage scores and part scores based on individual passage in a test on parts of speech sampled in the MCC format, and on grammatical and semantic constraints represented by a deletion (variations in the complexity and size of the informational unit that can be effectively processed by a reader.) These subscores, as well as passage cutoffs that define the reader's ability to read passages for a given purpose and within a given range of readability levels, enormously increase the diagnostic value of the test scores by providing the user with important clues to the reader's problems in the syntactic and semantic components of language.
Automated generation of items and tests	The objective, generative nature of both the item and test format makes possible the automation of item construction and test assembly and printing. This makes for both speed and economy, which in turn will permit the use of more complex but more useful evaluation designs in the schools.
Flexible resource	The cloze passages are part of a flexible test development resource. Instead of providing a set of fixed tests, the MCC corpus offers a collection of finely graded test passages allowing rapid and economical assembly of large numbers of special purpose tests to fit a variety of evaluation needs. This feature is especially important in situations where repeated diagnosis of skills and the measurement of growth in a narrow range of difficulty is desirable.

Objective Item Format

The objectivity or reproducibility of test construction procedures has, as noted previously, been a major criticism of norm-referenced measures of reading comprehension and this disadvantage was also evident

in the item development procedures in the basic competency model. These measures lack an explicit theory of what is meant by comprehension or basic reading competence and also objective or reproducible procedures for deriving items from theory. The development of the MCC format has involved a constant and, to date, largely successful effort to improve and maintain objectivity. Objectivity here, as noted, is important because, given other conditions, it enhances the possibility of repeatedly generating a test that incorporates an unbiased sampling of the content and behaviors of the universe of interest. The presumed unbiased nature of the test is, furthermore, traceable. Others interested in the operations defined by the test may generate similar or comparable tests, or the test may be generalized to other relevant behavioral domains in the course of extending or studing the underlying construct.

At latest study, the MCC format seems to offer the potential of being objectively reproducible. Several modifications in procedure and format now under consideration will reduce potential biases in passage selection that may have resulted from earlier, unnecessarily rigid limitations on passage length; the use of titles on passages; and possibly insufficient unmutilated context at the beginnings and ends of passages.

The current rule biased procedure for conversion of passages to the MCC format seems to offer additional potential for computerization, thus further approaching the possibility of reducing the item form to an algorithm. This computerization would presumably string together separate programs for readability analysis, conversion of passages to the mutilated format, and the generation and assignment of distractors from the word lists.

Unidimensionality

The MCC corpus constitutes a procedure for measuring a unidimensional construct called literal comprehension, a traditionally recognized reading skill that has never been adequately defined nor referenced to theories of language or reading. The separation here of this concept of literal comprehension from the global concept, comprehension, is based on the assumption that there are "degrees" of understanding or "levels" of meaning which may obtain between reader and text. Hence, literal comprehension is the apprehension of literal meaning ("explicit" or "clearly implied" meaning in teacher parlance).

Literal meaning is largely sentential meaning, and its technical definition relies primarily on work in single-sentence grammars (Chomsky, 1965; Katz, 1972). Thus, literal comprehension is defined as the apprehension of "the grammatical and semantic relations which obtain within and among the sentences of the discourse" (Katz and Fodor, 1963). The grammatical relations within the sentence which are crucial to its comprehension include subject-of, main-verb-of, object-of, and modifer-of (Chomsky, 1965). The grammatical relations among sentences referred to include explicit and implied connectors, anaphora, and ellipsis. Semantic relations include reference (Fodor, 1971) and categorical relations among words, etc. (Katz, 1972; Kintsch, 1972).

Traditional item types encompassed by such a definition include the wh-item (Bormuth, 1970), items measuring the identification of word meanings in phrases or sentential contexts, "detail" and "fact" questions, paraphrase and summary items, best title questions, explicit main idea questions, and questions about implications which hold within the test passage. Many variations on the cloze procedure also assess literal comprehension as defined.

The educational significance of the term literal comprehension lies in the proposition that this is the level of comprehension which is essential to an enormous variety of meaningful life roles, such as reading recipes and newspapers and reacting appropriately to warnings, consumer information, and so on. Literal comprehension also appears to be prerequisite to the more extensive processing of text referred to as inferential comprehension, critical and evaluative reading, and so on, which may involve higher order reasoning skills.

Research to date with the MCC corpus in a student population of 5,000 grade 1-9 students, including various classifications of retarded and disadvantaged readers, provides considerable support for the unidimensionality of the construct of literal comprehension and its relative independence from IQ (O'Reilly et al., 1976; O'Reilly and Streeter, 1977). In this research the r between tests in the MCC format and a group measure of verbal IQ ranged from a low of .35 among first graders to a high of .62 among sixth graders (N=5,000). In contrast, standardized measures of reading comprehension typically correlate in the 70's and 80's with IQ in grades 1-3 and in the 80's and 90's in grade 4 and above (Singer, 1973).

Passage Dependency

The validity of several well-known tests of reading comprehension has been seriously challenged on the issue of passage dependency--the tendency for students to obtain scores well above chance without reading the test passages. Some authors have proposed that this issue be handled by redefining comprehension as information gain (Bormuth,

1970) or as a residual gain (Rankin and Dale, 1969) rather than by making better tests. These procedures attempt to remove from the test score the influences of specific and general knowledge and various other test-taking strategies that operate on test questions independent of the test passages.

Though one might wish to define new learning in this way, i.e., as information gain, redefining comprehension in such terms seems to create new problems. One has only to consider for a moment the virtual impossibility of distinguishing between that part of the test score that represents prior knowledge of the reader and the part which represents new learning. How does one ask a question about a passage in the pretest situation which does not involve the interplay of old and new information or knowledge?

It does not seem desirable, necessary, or possible to attempt to remove the influences of specific, general, or idiosyncratic knowledge from comprehension test scores. According to the theory underlying the test, all meaning exists in the reader, and reading or comprehending necessarily involves bringing such meaning to bear on comprehending a passage. The best available solution to the problem of passage dependency, therefore, involves eliminating as much as possible the influences of additional meaning or other irrelevant informational cues offered by the test situation. Such influences are necessarily present in the typical multiple-choice questions in tests of comprehension (e.g., idiosyncratic meanings are introduced by the ways in which the test-writer interprets a given passage), but it would be difficult or

impossible to eliminate them without also seriously compromising the measurement of comprehension.

The multiple-choice cloze format appears designed to avoid introducing into the test situation idiosyncratic meaning and irrelevant cues that both affect the test score and interfere with the reader's attempt to process independently the information in the test passages. The procedure for generating distractors that compete grammatically but not semantically is specifically designed to limit, insofar as possible, the context for interpretation of the test passage. Prior experimentation with the MCC format indicated that the use of semantically interfering distractors would have the effect of introducing a very difficult vocabulary element into the test with the further effect of increasing the correlation of the test with IQ (Cranney, 1972). The distractor design is also intended to enhance test validity by maintaining the passage dependency of the test. The use of grammatically equivalent responses in the MCC item format should function to eliminate the use of grammatical cues as the only basis for choosing among distractors. Similarly, the inclusion of competing content words among certain groups of distractors should tend to eliminate discontinuities in content as a basis for choosing among distractors. In addition, the design of the MCC test format is basically coherent with the ongoing act of comprehension, which is characteristically rapid and focused on the processing of relatively large information units. The structure of the passages in a given test is unchanged by the multiple-choice cloze format. The act of

responding to the word choices remains focused on the meaning of the passage, interrupted only by periodic deletions. In this act, the reader predicts the words that correctly replace these deletions to complete the intended meaning by drawing upon his own relevant cognitive structures, unique or otherwise, and not, as may be the case with other formats, those in the mind of a test-maker.

Domain-Referencing and Scaling

The passages in the MCC corpus, like those in other measures of basic reading competence, were systematically drawn from clearly defined and relevant domains of written discourse. The selection of relevant domains was aided by reference to the Hansen and Hesse effort in Madison, Wisconsin (1972) to build a domain-referenced test based on the standard cloze. In that effort parents and teachers identified relevant domains of written discourse on the basis of frequency of use or importance in the school and community for students in grades 4 through 12. The present effort has improved on the relevance and specificity of these domains by extending the domains and levels to texts used in grade 1 and also by incorporating readability and potentially other linguistic features as additional defining characteristics of texts and other relevant discourse.

Using the domain-referenced model as a basis for assembling the variety of available MCC passages is a deliberate attempt to maximize the relationship between the test situation and a variety of important school and extra-school reading situations. Maximizing this relationship should, in turn, facilitate the production of tests of reading

comprehension that are maximally sensitive to some of the most important outcomes of reading instruction. Such tests would be designed so that the content sampled in the test is relevant to and reflects the changing nature of the experiences of readers with written discourse over time. A suitable test for a beginning reader would thus present the testee with a systematic selection of passages sampled from the distribution of readability of texts available to first graders, as shown in Figure 2. Repeated administration of comparable tests would then demonstrate the reader's increasing ability to comprehend available texts and would provide a continuing basis for guiding the selection of suitable texts.

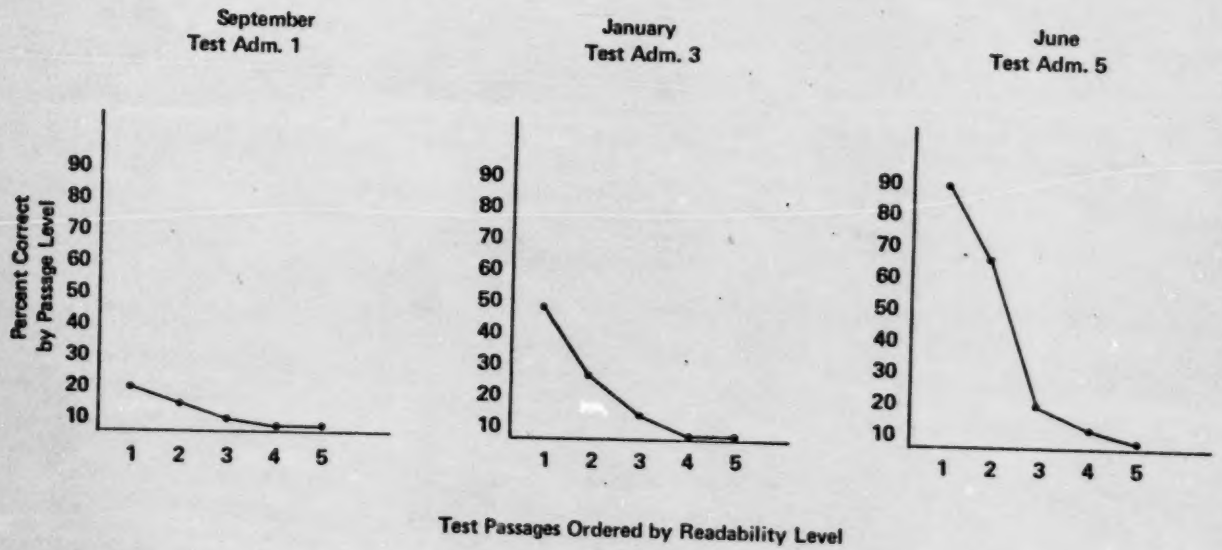

Figure 2 Progress of a beginning reader in comprenhension as measured by passages ordered on readability and sampled from first grade texts.

An appropriately designed test of literal comprehension will thus generate a series of scores or score distributions over time which will reflect the increasing ability of a given student or population to apprehend the literal meaning of increasingly difficult passages and more diverse domains. The nature of these performance distributions is a function of the time between test administrations, the degree of relevance of the content and the difficulty levels of the passages selected for the test, the ability of the students, and the degree of learning that occurs in the time framework of the test. The ability of such a test to measure certain broad effects of reading instruction and experience is theoretically maximized when the test is designed to generate the types of scores shown in Figure 2 for each level of a given instructional system.

The ability to utilize data, such as that illustrated in Figure 2 or a distribution of scores derived from a similar testing arrangement, in selecting texts or other discourse for individuals and groups is further dependent upon establishing a formal link between the scale used to calibrate passages and that used to calibrate the comprehensibility of discourse domains. Currently, the comprehensibility of the passage corpus has been scaled through the application of the Spache and Dale Chall formulas. Ultimately, the test passages must also be scaled based on human performance and all test passages referenced to a single overall scale. Criteria of passage performance must also be established that are consistently applicable, if possible, across the total range of passage difficulty. With this proposed scale and the

associated passage performance criteria (e.g., the passage is at the student's frustration level . . . the student's instructional level), it becomes possible to express, within error of measurement, the levels of passage difficulty that are suitable for a given reader for certain broad purposes, such as independent reading.

Given that the test passages are simultaneously scaled using a methodology which can be directly applied to written discourse, such as a readability formula, and this scale accurately reflects the same linguistic dimensions that contribute to the scaling based on human performance (i.e., the two scales rank the test passages similarly), then the stage is set for linking or generalizing test scores to specific domains of materials. What is then required is that the readability formula be systematically applied to various domains of relevant reading materials to define their respective distributions of comprehensibility. Individual test scores may then be referenced to these distributions, or distributions of test scores may be superimposed on the distributions of comprehensibility of various domains of texts available by grade level, as illustrated in Figure 3.

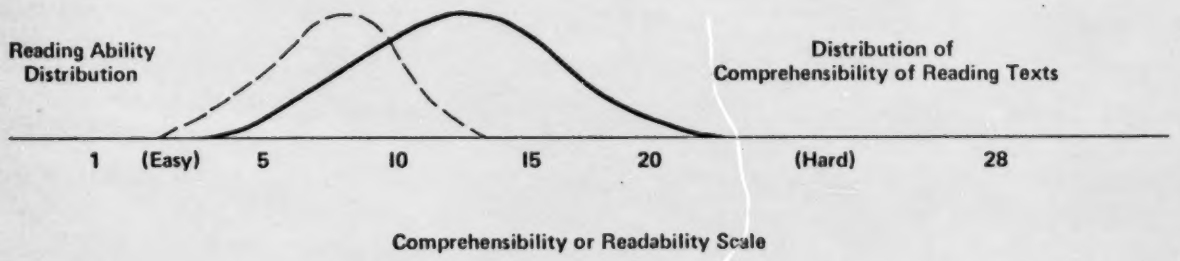

Figure 3 Distribution of comprehensibility of reading texts in grade 2 compared to the distribution of reading ability in grade 2

The feasibility of achieving these features of the model was recently subjected to a preliminary test in New York State[5] in a study involving the calibration of some 214 MCC passages using both the Rasch measurement model (Wright 1968) and the Spache-Dale Chall continuum. Some 10,000 students in grades 1-12 each received one of several MCC test forms in a complex linking design that permitted the calibration of the total passage pool along a single Rasch continuum. Spache and Dale Chall values were also available on the same passages and the two sets of scale values were subsequently correlated. Preliminary results show the correlations between the two scales to be: .94 for grades 1-3, .94 for grades 4-6, and .87 for grades 7-9. Summary statistics for these analyses are presented in Table 2.

These data indicate that the readability values of textual materials available to the student population can be predicted with substantial accuracy from item difficulties based on performance (within a grade level, Rasch calibrations of MCC passages and traditional item difficulties on those passages are correlated in the 90's). Given that the reliabilities of individual passage scores can be raised somewhat, it appears that the test model can be used to select materials that can be comprehended at the literal level by individual students

5 This study was organized by the authors while still at the New York State Education Department and in conjunction with Steven Kidder, who will report in more detail on these findings at this AERA session.

Table 2

Summary of Statistics and Regressions of Rasch Calibration(Y) on Readability Scores(X)

	1 Spache	2 Dale-Chall	3 Spache and Dale-Chall
Mean X	2.59	6.47	5.54
Mean Y	121.29	303.66	259.89
SDx	0.77	1.07	1.95
SDy	15.58	72.36	100.64
SEM_x	0.11	0.09	0.14
SEM_y	2.25	5.87	7.12
N[a]	48	152	200
r_{xy}	0.91	0.98	0.97
linear regression:			
a (intercept)	74.02	-124.70	-18.77
b (regression weight)	18.27	66.163	50.29

a N's reflect number of passages.

and groups of students. Internal reliability for the total test, as shown in Table 3, is very high, but individual passage scores are generally less reliable than is required for individual prediction. Acceptable reliability can very likely be achieved by increasing passage length to 12-15 items and/or by increasing the consistency of the pattern of deletions through changes in the item writing rules. Both possibilities are presently being investigated.

Table 3

Kuder-Richardson Formula 20 Reliability Coefficients for MCC Passages in a Second Grade Sample (N = 4000)

	Passage No.								
	1	2	3	4	5	6	7	8	Total
KR-20	.59	.70	.75	.83	.85	.86	.88	.91	.96
No. Items	3	3	5	5	10	10	10	10	56

Diagnostic Part Scores/Passage Performance Criteria

The cloze test is intended to be presented to the testee as a series of passages finely graded on readability. The user then identifies the cutoffs indicating the student's point of illiteracy, instructional reading level, and independent reading level by examining the response pattern at each level of passage readability. Visual fitting of curves, using wh-questions as a criterion (O'Reilly et al., 1976) indicates the respective cutoffs appear to be 0-3, 4-7, and 8-10 items correct for 10 item passages. However, the total pattern of

responses across passages must be examined to ensure accuracy in drawing the cutoffs because reversals do occasionally occur. This procedure has proven to be extremely useful in identifying student's reading levels.

Given the highest readability level at which the student indicates an ability to comprehend independently, a reading teacher might (1) assign students to read texts of comparable content and difficulty for literal meaning without instructional assistance, (2) assign students to read slightly more difficult texts for literal meaning with instructional assistance, or (3) assign students to read texts of comparable difficulty critically or for inferred meaning with instructional assistance. Item 1 above assumes that the MCC task is a surrogate for a host of other tasks considered to access literal meaning (e.g., "detail" questions). Item 2 assumes that students can learn to read more difficult texts by gradual exposure to them with instructional assistance. Item 3 assumes that literal comprehension ability does not generalize to inferential skills and critical reading ability but does create a solid foundation for instruction in those higher level reading skills.

In more recent work with the MCC passages, a series of experimental subscores have been developed to give the reading teacher yet more detailed diagnostic information on the student's understanding of syntactic and semantic features of a set of test passages. Here the focus is on interpreting the student's failure to select the correct words for specific categories of deletions in a set of MCC passages based on the notion that such failures are attributable to one or both

of two general causes: (1) Either the vocabulary of the exercises exceeds students' working word knowledge, or (2) the sentences are too long and/or complex for the students to "parse" (i.e., identify the subject, main verb, etc). More task-specifically, poor performance indicates an inability to use contextual clues to recognize missing words in written discourse among sets of distractors which match the missing words in part of speech and vocabulary level. Beyond that, subscores are required for score interpretation.

Part of speech. There are two kinds of subscores on MCC exercises: (1) part of speech and (2) contextual constraint. Because there are so few items representing a given subscore (e.g., noun scores) available on a given passage, subscores are derived from the entire set of items of that kind available on a test form. Noun subscores, for instance, would reveal the percent of noun deletions recognized correctly out of the total number of nouns deleted on the entire test form. The point of interest is significant discrepancies in ability to recognize missing words across part-of-speech categories. Percent correct on noun scores, for example, might be 60,while percent correct on verb scores might be 30. Such a configuration would indicate difficulty in using contextual clues to recognize missing verbs in written discourse and would ascribe literal comprehension difficulties to that particular problem.

Each part of speech subscore is also capable of a more refined interpretation. Noun scores, in general, indicate knowledge of the subject matter of the text from which the passage was excerpted and, more particularly, recognition of the topical content of the excerpted

passage itself.

But nouns function differently in different syntactic positions in the paragraphs (e.g., subject, direct object, etc.), so low noun scores might indicate an inability to interpret nouns in particular positions in sentences. Literal comprehension problems, then, might be identified as lack of experience in, say, interpreting nouns in direct object position. Such a problem, so identified, is relatively easily remedied by instruction.

On the other hand, noun scores might be high in comparison to verb scores. If noun scores in general indicate recognition of topical content, then verb scores indicate recognition of what is predicated of that topical content. If nouns reveal that the passage is "about" Dick and Jane, then the verbs tell what Dick did to Jane, a matter of no small consequence in comprehension. Verbs, however, also occur in variations as to tense, number, etc., so verb scores might reveal problems with inflectional endings, irregular past tenses, and so forth. In addition, verbs occur in positions other than main verb. Infinitive forms might be one locus of comprehension problems, or verbs in complement position may be a source of difficulty. Given such a detailed identification of the breakdown in comprehension, remedial action is specific and often available in standard instructional programs.

Adjectives are theoretically more complex. They are often verb forms (e.g., past participles) in adjectival position (i.e., modifying nouns, etc.). Adjective scores, therefore, should always be analyzed with attention to concommitant problems with verbs. Secondly,

adjectives are often considered embedded sentences--e.g., the two sentences, The kid is benign and The kid is fat, may be reduced to one sentence by embedding the second sentence in the first as as adjective: The fat kid is benign. Students having difficulty with adjectives may lack experience with verbs in adjectival position, or they may lack experience in reduction of sentences to adjectives.

The adverb score represents a yet higher order of complexity than the adjective score. According to Bever (1970 p.), adverbs signal complex relationships in the discourse and, to be interpretable, occur in a prescribed order within the sentence:

direction
manner
place
duration
frequency
time
purpose

which is attested to by attempts to rearrange the order of the following sentence and maintain comprehensibility:

Georgeala Cough rode north fast in the park briefly

often yesterday for fun.

The adverb score is thus potentially indicative of the student's failure to perceive the various types of relationships signalled by adverbs as well as by adverbial phrases and clauses within the sentences of a discourse. One implication for the teacher is to work with the student on a variety of visually and orally presented sentences containing appropriate examples.

Contextual constraint. As in any cloze procedure, the choice of words to fill in blanks in MCC exercises is constrained by the verbal

context in which the blank space occurs. There are essentially two kinds of constraints operating on the choice of missing words in connected discourse: (1) the nature of the language itself--its structural properties and meaningful relationships with cognitive experience--and (2) patterns of language use within a particular speech community (e.g., academia, journalism, dialects). Patterns of use constrain the appropriateness of a word in a given context (e.g., "slang" in a funeral oration in middle-class society) and affect reader/writer roles, relationships, the tone of the message, etc.

While such patterns of use affect the choice of the missing words in a free-response (fill-in-the-blank) cloze procedure, the distractor-selection procedures in the MCC exercises effectively eliminate appropriate/inappropriate as a criterion for choosing the correct answer. (The MCC item form is not sensitive to tone, culturally biased, etc.) Instead, the choice of words to fit in blank spaces in the MCC exercises is constrained primarily by the nature of the language itself. Thus the criterion for choosing the missing word from among a set of responses in MCC exercises is acceptable/unacceptable (grammatically), makes sense/doesn't make sense, rather than appropriate/inappropriate, which is a more refined judgment than literal comprehension demands. Thus the constraints affecting the choice of words in the MCC exercises are grammatical and semantic; only the correct answer makes grammatical and semantic sense in the passage.

Beyond those general observations, however, the <u>extensiveness</u> of the grammatical and semantic constraints affecting the recognition of the missing word in a given passage varies widely. In other words,

some items can be answered correctly by reading less than a full sentence in the passage (intrasentential constraint); others require reading the entire sentence (sentential constraint); and some items demand information from beyond the sentence in which the item occurs (intersentential constraint).

Consider the following item where only two words of context preceding the deletion are retained:

. . . mothers and ____4____

1. covers
2. cards
3. keepers
4. fathers

A high proportion of students can choose the correct answer--"fathers"--with no more context than the two words preceding the deletion. The rest of the sentence increases the certainty that "fathers" is the correct answer, but two words is nevertheless sufficient information: "He could hear the mothers and ____4____ as they talked." Hence the constraint is mostly local or intrasentential. (Locally constrained items are often over-worked phrases requiring little more than word-association skills to recognize the right answer.)

Other items are primarily sensitive to sentential contexts; that is, scores increase as context is added up to the full sentence in which the deletion occurs but change little as context beyond the sentence is added. Constraint on such items is primarily sentential. An example follows:

Debbie said, "Come here, David. I will help you clean your shoes."
"Let me do it," ___3___ David.

1. looked
2. saw
3. said

"You can clean your shoes. But you do not have to clean my shoes. I do not want help."

Finally, some items require information from beyond the sentence in order to recognize the right answer. In this case, constraint is intersentential. Given only the sentence, "Puff wants a ___2___."

1. friend
2. ride
3. step

for example, it is difficult to choose the correct answer, "ride." But the following sentences supply the necessary information: "Sally said, 'No, no, Puff! You can not ride in that car . . . .'"

The foregoing part scores represent an innovative approach to diagnosing reading problems in comprehension, an approach which is based on identifying the specific semantic and syntactic problems the students may have in processing a sample text. Extant comprehension tests do not specify in sufficient detail the kinds of conceptual and language problems that are responsible for the student's inability to comprehend a class of textual materials. A separate vocabulary score, for example, does not identify the specific vocabulary that causes the student to fail to apprehend the major ideas and concepts that run through the text. Comprehension subscores, such as relationships and generalizations fail to signal anything about the underlying language

structure of the text that may contribute to the student's reading problems. Relatively recent research on reading and language (Sitko and Semmel, 1972; Blake, 1975; Vogel, 1975; Weinstein and Rabinovitch; Guthrie, 1973), however, clearly indicates that both syntactic and semantic features of language are strongly implicated in comprehension failures in children.

This approach to reading diagnosis may also be closely integrated with the readability analysis methodology discussed previously. The extension of the MCC format to the generation of experimental part scores proposed here has its counterpart in the application of readability procedures to various textual domains. Such an analysis might, for example, parallel the MCC part scores by listing the unique vocabulary of texts from level to level as well as other factors describing features of the language affecting comprehension.

Conclusion

This discussion began with recognition of the need for new measures of reading performance that would allow users to confidently judge how well readers would perform in a variety of meaningful reading situations. New models of reading assessment directed at this need, called basic competency tests, tests of functional literacy, and so, were reviewed as a whole and found to offer only very modest capacity for assessing basic reading competence. Essentially the new tests primarily offer the feature of domain-referencing, but otherwise share many of the defects that have limited the utility of norm-referenced tests as basic measures of reading outcomes in the school situation.

An alternative model, based on a modification of the cloze technique, was offered as an approach that comes closer to fulfilling

needs for a measure of basic reading competence across the levels of the school curriculum. The MCC format shares the feature of domain-referencing with the new basic competency tests, but otherwise offers several additional features that further enhance its utility as a measure of a basic level of reading comprehension. Items in the test are objectively developed and derived from a theoretical framework, thus providing guidelines for exploring and interpreting the meaning of resultant test scores. Current developments on the test strongly indicate the feasibility of using a test score to identify the kinds of materials that readers can comprehend at the literal level for specific purposes. And, the test appears to offer a diagnostic capability with fairly clear-cut instructional implications, although this feature requires considerable further exploration in research and practice.

New work underway now in the test promises to deliver much more useful information about the domains of materials that are relevant to the various age-graded populations of readers. For example, the authors are currently planning the development of a new readability analysis procedure that will deliver additional data on the domains of discourse of interest, such as basal readers, texts in the content areas, and other relevant materials. Over a three-year period, some 5,000 texts and other sources will be submitted to analysis via this program and the resultant domain distribution of readability defined. As one outcome of this procedure, it is expected that new standards of basic reading performance can be set for students in grades 1-12. These standards will be defined around the domain distributions of readability that

become available on various types of reading materials and the relation of these distributions to the reading requirements of various life roles. For example, using this approach, it will be possible to define in terms of readability and other discourse variables the basic reading requirements for a beginning second grader or a graduating senior.

Additional work is planned on the yet tentative passage performance criteria used with the MCC passages, as well as in the underlying meaning of the test scores. In the former study, multiple item forms will be written for 4th and 6th grade passages and used as a basis to form three groups representing frustration, instructional, and independent reading levels. Cutting scores which optimally discriminate among the three groups in comparable MCC passages will then be defined. In the latter study planned for this Spring, children's MCC scores will be examined in relation to their ability to infer the underlying propositions on the same and similar passages using the methodology employed by Frederiksen (1976).

The MCC as a model for assessing basic reading competence is not applicable to some classes of reading stimuli and situations. However, it seems a reasonable supposition that literal comprehension, as defined, is the basic skill involved in a large percentage of the basic reading situations identified in studies of basic literacy, such as the ETS research previously referred to.

REFERENCES

Anderson, R. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170

Bever, T. G. The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), Cognition and the development of language. New York: John Wiley, 1970, 279-362.

Blake, K. Type of transformation and retarded and normal pupils' comprehension of sentence material. Journal of Research and Development, 1975, 8, 89-94.

Bormuth, J. On the theory of achievement test items. With an appendix by P. Menzel: On the linguistic bases of the theory of writing items. Chicago: University of Chicago Press, 1970.

Bormuth, J. Development of standards of readability: Toward a rational criterion of passage performance. Final Report. (U. S. Office of Education: Project No. 9-0237.) Chicago: University of Chicago, 1971. (ERIC Document Reproduction Service No. ED 054 233)

Carroll, J., Davies, P., & Richman, B. The American Heritage word frequency book. New York: Houghton-Mifflin, 1971.

Carver, R. P. Reading as reasoning: implications for measurement. In W. J. MacGinitie (Ed.), Assessment problems in reading. Newark, Delaware: International Reading Association, 1973.

Chomsky, N. Aspects of the theory of syntax. Cambridge, Mass.: MIT Press, 1965.

Cranney, A. The construction of two types of cloze reading tests for college students. Journal of Reading Behavior, 1972, 5, 60-64.

Fillenbaum, S., Jones, L., & Rapoport, A. The predictability of words and their grammatical classes as a function of rate of deletion from a speech transcript. Journal of Verbal Learning and Verbal Behavior, 1963, 2, 186-194.

Fodor, J. A. The ontogenesis of the problem of reference: A review of some theories of linguistic symbols. In C. A. Reed (Ed.), The learning of language. New York: Appleton-Century-Crofts, 1971, 333-362.

Frederiksen, C. H. Semantic processing units in processing text. National Institute of Education, 1200 19th Street, N.W., Washington, D. C. 20208, 1976.

Hansen, L. H., & Hesse, K. D. An interim report of results of the pilot assessment of reading literary. Madison, Wisconsin: Madison Public Schools, the Office of Research and Testing in the Department of Curriculum Development, 1972.

Harnischfeger, A. & Wiley, D. W. The marrow of achievement test score declines. Educational Technology, June, 1976, 16, 5-14.

Harris, L., et al. Survival literacy study. New York, 1970. ED 069813

Harris, A., & Jacobson, M. Basic elementary reading vocabularies. New York: Macmillan, 1972.

Hively, W. Domain-referenced testing: Part one, basic ideas, introduction to domain-referenced testing. Educational Technology, 1974, 14, (6), 5-10.

Katz, J. Semantic theory. New York: Harper & Row, 1972.

Katz, J., & Fodor, J. The structure of a semantic theory. Language, 1963, 39, 170-210.

Kintsch, W. Notes on the structure of semantic memory. In E. Tulving and W. Donaldson (Eds.), Organization of memory. New York: Academic Press, 1972, 247-308.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Murphy, R. T. Adult functional reading study. Final Report. (Project J: Targetted Research and Development Reading Program Objective, Subparts 1, 2, and 3, Project No. 0-9004.) Princeton, New Jersey: Educational Testing Service, 1973.

Nafzinger, D. N., Thompson, B. R., Hiscox, M. D., & Owen, T. R. Tests of functional adult literary: An evaluation of currently available instruments. Portland, Oregon: Northwest Educational Laboratory, 1976.

National Assessment of Educational Progress, Reading and Literature: General Information Yearbook. Report #02-61Y. Washington, D. C.: U. S. Government Printing Office, 1972.

National Institute of Education. Assessment of reading comprehension, National conference on studies in reading, 1200 19th street, N.W., Washington, D. C., 1975.

Northcutt, N., Kelso, C., & Barron, W. E. Adult functional competency in Texas. Austin, Texas: University of Texas Press, 1975.

O'Reilly, R. P., Schuder, R. T., Kidder, S. J., Salter, R. & Haylord, P. D. A report on the validation and refinement of measures of literal comprehension in reading for usein policy research and classroom management. Albany, New York: The University of the State of New York, The State Education Department, Division of Research, December, 1975.

O'Reilly, R. P., & Streeter, S. J. Reports on the development and validation of a system for measuring literal comprehension in a multiple-choice cloze format: Preliminary factor-analytic results. Journal of Reading Behavior, 1977 (In Press).

Rankin, E. F. & Dale, L. H. Cloze residual gain--a technique for measuring learning through reading. In G. B. Schich and M. M. May (Eds.), The psychology of reading behavior, Eighteenth Yearbook of the National Reading Conference. Milwaukee, Wisconsin: National Reading Conference, 1969.

Schuder, R. T., Kidder, S. J., and O'Reilly, R. P. Paper presented at the annual convention of the American Educational Research Association, San Francisco, California, 1976.

Singer, H. IQ is and is not related to reading. Paper presented at the annual Convention of the International Reading Association, Denver, Colorado, May 6, 1973.

Sitko, M. C., & Semmel, M. I. The effect of phrasal cueing on free recall of EMR and non-retarded children. American Educational Research Journal, 1972, 9, 217-229.

Taylor, S., Frackenpohl, H., & White, C. A revised core vocabulary: A basic vocabulary for grades 1-8, an advanced vocabulary for grades 9-13 (Research and Information Bulletin No. 5). Huntington, New York: Educational Developmental Laboratories, March, 1969.

Tuinman, J. Determining the passage dependency of comprehension questions in 5 major tests. Reading Research Quarterly, 1973-1974, 9, 206-223.

Vogel, S. Syntactic abilities in normal and dyslexic children. Baltimore, Maryland: University Park Press, 1975.

Weinstein, R., & Rabinovitch, M. Sentence structure and retention in good and poor readers. Journal Educational Psychology, 1971, 62, 25-30.

Wright, B. D. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1968.